

# iBus Data Cleansing

## for Quality of Service Indicators (QSI)

v2.1 01.04.2015

---

The purpose of this document is to assist Bus Operators in understanding the iBus QSI Data Cleansing and Exclusion processes.

This document replaces the 'iBus Data Cleansing' v2.0 published in December 2013.

The document removes reference to Hyperion Report R001 which no longer shows total excluded data.

The document is part of a set describing the QSI processes under iBus which also includes 'iBus Missing Data Mitigation for QSIs', 'iBus Data Aggregation for QSIs' and 'iBus QSI Statistics Explained'.

# Contents

---

1. Introduction .....	3
2. Stage I: Missing Data Mitigation.....	5
Overview.....	5
High Frequency Data Imputation.....	5
Low Frequency Data Removal .....	5
Technical Details .....	5
3. Stage II: Automatic Exclusion .....	6
Overview.....	6
High Frequency Automatic Exclusion.....	6
Low Frequency Automatic Exclusion.....	7
4. Stage III: Manual Exclusion.....	9
Overview.....	9
Unrepresentative Service .....	9
System Failure.....	11
5. Bus Operator Data Exclusion Request Process.....	13
6. Data and Process Integrity.....	14
7. Data Exclusion Codes.....	15
8. Hyperion Performance Reports.....	16
Overview.....	16
Unfinalised Data .....	16
100% Data Excluded .....	16
R001H/L 'Quality of Service Indicators (with Minimum Standards)' .....	16
R210 'Trip Observation Detail'.....	17

# 1. Introduction

---

- 1.1. The role of QSIs is the measurement of service reliability as perceived by the passenger and iBus provides the data from which QSIs are calculated. Operators are expected to take all reasonable measures to maintain reliability in the event of both foreseeable and unforeseeable disruptions.
- 1.2. iBus data provides comprehensive and continuous coverage and will capture conditions that are favourable to a route's performance along with those which are less so. Therefore 'cleansing' of data recorded at times of significant service disruption is not required simply to ensure the sample is representative of the passenger experience for a particular time period. Nevertheless, a number of data cleansing processes (both automatic and manual) are included within the iBus system to mitigate against:
  - 1.2.1. Missing data.
  - 1.2.2. Erroneous data where the error(s) cannot be satisfactorily corrected.
  - 1.2.3. Data that could distort the QSI results in terms of reflecting the passenger experience.
  - 1.2.4. Major service disruption.
- 1.3. Data will be excluded from the overall sample in stages with the minimum unit of exclusion being one hour for a particular route, direction and QSI point. This allows a very detailed level of data to be excluded and, therefore, increased accuracy in the reported results.
- 1.4. The first stage of 'data cleansing' is Missing Data Mitigation (data imputation/removal process) which occurs automatically and is designed to remove the impact of 'missed buses'. The techniques applied are dependant on QSI frequency status. See Section 2 for details.
- 1.5. The second stage is an 'automatic exclusion' process which replaces the previous 'legacy' exclusion process and is designed to remove unrepresentative data caused by major service disruption. The techniques applied are dependant on QSI frequency status. See Section 3 for details.
- 1.6. The above automatic processes may, at the discretion of London Buses, be supplemented by additional manual exclusions to remove other **data** errors in the system. Examples include 'false starts' and other vehicle navigation issues. See Section 4 for details.

- 1.7. All data is excluded at the discretion of London Buses and there shall be no right of appeal for service disruptions. However, London Buses acknowledge that there may be occasions where the processes outlined above may not identify and remove all erroneous data. A mechanism is in place whereby operators may submit a request for London Buses to review data that appears to be erroneous. See Section 5 for details.

## 2. Stage I: Missing Data Mitigation

---

### Overview

- 2.1. There are a number of technical reasons why some actual departure times at QSI points may not be recorded in iBus as observed data. Left unchanged this 'missing data' could have a negative impact on QSI results. The data imputation/removal process is designed to mitigate for missed buses.
- 2.2. The data imputation/removal process takes place weekly immediately after the MTV Mileage Coding Finalisation process.
- 2.3. The techniques applied to mitigate the impact of missed buses are different for high and low frequency routes.
  - 2.3.1. **High Frequency** routes use a method of data imputation.
  - 2.3.2. **Low Frequency** routes use a method of data removal.

### High Frequency Data Imputation

- 2.4. For a **high frequency** route, when a missing trip record is coded via the MTV with an imputable cause code, iBus will aim to impute an observed time for any QSI points within the missing section. To do this accurately, iBus needs a certain amount of observed data from other buses within a specified time period (time periods are mostly 3 hour blocks but can be 2 or 5 hours). If sufficient data is not available for a particular route/QSI point/direction, all data within the time period for that route/QSI point/direction will be excluded from the QSI calculations. This is an automatic process.

### Low Frequency Data Removal

- 2.5. For a **low frequency** route, when a missing trip record is coded via the MTV with an imputable cause code, iBus will ignore such trips for QSI points within the missing section in the 'linking' process that produces the punctuality statistics and also in the count of expected buses. Consequently no data exclusions arise from this process. This is an automatic process.

### Technical Details

- 2.6. For full technical details refer to the supporting document entitled 'iBus Missing Data Mitigation for QSIs'.

## 3. Stage II: Automatic Exclusion

---

### Overview

- 3.1. Following data imputation/removal, London Buses will apply an automatic data exclusion process.
- 3.2. Automatic exclusions are designed to remove unrepresentative data arising from exceptional disruption to services caused by both planned and unplanned events. Such 'events' include but are not limited to:
  - 3.2.1. Road traffic collisions, roadworks and short-term road and lane closures.
  - 3.2.2. Planned and unplanned incidents, such as demonstrations, sporting fixtures, adverse weather, security incidents, terrorist attacks, police incidents, power outages, traffic signal failures, natural disasters, fires and industrial action.
  - 3.2.3. Incidents impacting on other transport modes such as road, rail or ferry services.
- 3.3. The proportion of 'observed buses' against 'expected buses' is calculated for all hourly totals. Where this proportion falls below a certain threshold the hour is marked as excluded, on the assumption that the shortage of observed buses in these hours has been caused by data errors or wholly unrepresentative operating conditions.
- 3.4. The techniques applied for excluding hours are different for High and Low Frequency routes.
  - 3.4.1. **High Frequency** routes use calculations based on the proportion of observed buses (after imputation) against expected buses.
  - 3.4.2. **Low Frequency** routes use a calculation based on the number of time linked buses (the sum of 'on-time', 'early' and 'late' buses).

### High Frequency Automatic Exclusion

- 3.5. For a **high frequency** route, the proportion of observed buses (after imputation) against expected buses is calculated for all hourly totals. Where this proportion falls below 30% for any route/direction/QSI point, two rules are activated.
  - 3.5.1. Rule 1: the hour is automatically excluded.
  - 3.5.2. Rule 2: any hour immediately following an hour excluded under rule 1 at a given location and in the same direction is automatically excluded. The following hour must be +1 hour from the 'rule 1' excluded hour.
    - **Example 1** below shows automatic exclusions for a high frequency route.

## Example 1

Route	Hour	QSI Point	Direction	Scheduled	Observed	Excluded	Observed / Scheduled %	Comments
1	01/04/12 10:00	QSIABC Q	1	8	7	0	88%	
1	01/04/12 11:00	QSIABC Q	1	7	7	0	100%	
1	01/04/12 12:00	QSIABC Q	1	10	3	0	30%	Do not exclude, as % observed is not <30%
1	01/04/12 13:00	QSIABC Q	1	9	1	6	11%	Hour is already excluded via Data Imputation Process
1	01/04/12 14:00	QSIABC Q	1	10	2	4	20%	Automatically exclude as % observed is <=30% (rule 1)
1	01/04/12 15:00	QSIABC Q	1	9	2	4	22%	Automatically exclude as % observed is <=30% (rule 1)
1	01/04/12 16:00	QSIABC Q	1	10	6	4	60%	Automatically exclude as preceding hour has <=30% observed data (rule 2)
1	01/04/12 17:00	QSIABC Q	1	8	1	4	13%	Automatically exclude as % observed is <=30% (rule 1)
1	01/04/12 18:00	QSIABC Q	1	6	1	6	17%	Hour is already excluded via Data Imputation Process
1	01/04/12 19:00	QSIABC Q	1	6	4	4	67%	Automatically exclude as preceding hour has <=30% observed data (rule 2)
1	01/04/12 20:00	QSIABC Q	1	5	3	0	60%	
1	01/04/12 21:00	QSIABC Q	1	5	4	0	80%	
1	01/04/12 22:00	QSIABC Q	1	5	2	0	40%	
1	01/04/12 23:00	QSIABC Q	1	5	5	0	100%	
1	02/04/12 00:00	QSIABC Q	1	2	0	4	0%	Automatically exclude as % observed is <=30% (rule 1)
1	02/04/12 05:00	QSIABC Q	1	4	3	0	75%	Do not exclude, the previous hour is 5 hours before and is not the preceding hour.
1	02/04/12 06:00	QSIABC Q	1	6	5	0	83%	
1	02/04/12 07:00	QSIABC Q	1	8	7	0	88%	

- 3.6. Excluding any hour immediately following a period where the proportion of observed buses against expected buses falls below 30% for high frequency routes as outlined above provides further protection in terms of allowing operators a reasonable “grace” period to recover from the disruption.

### Low Frequency Automatic Exclusion

- 3.7. For a **low frequency** route the automatic exclusion process is based on buses linked using the ‘linking by time’ process (see the supporting document ‘QSI Statistics Explained’ for further details). The hour will be excluded for a route/direction/QSI point where the number of time linked buses (the sum of ‘on-time’, ‘early’ and ‘late’ buses) is:
- 3.7.1. 0 or 1, for hour blocks where the number of expected buses (after imputation) is 3 or above.
- 3.7.2. 0, for hour blocks where the number of expected buses (after imputation) is 2.
- **Example 2** below shows automatic exclusions for a low frequency route.

## Example 2

Route	Hour	QSI Point	Direction	Scheduled	Observed	Excluded	Comments
2	01/04/12 10:00	QSIXYZ Q	1	2	1	0	
2	01/04/12 11:00	QSIXYZ Q	1	2	0	4	Automatically exclude as no. of linked buses = 0
2	01/04/12 12:00	QSIXYZ Q	1	3	2	0	
2	01/04/12 13:00	QSIXYZ Q	1	3	1	4	Automatically exclude as no. of linked buses <2
2	01/04/12 14:00	QSIXYZ Q	1	4	4	0	
2	01/04/12 15:00	QSIXYZ Q	1	6	1	4	Automatically exclude as no. of linked buses <2
2	01/04/12 16:00	QSIXYZ Q	1	4	2	0	
2	01/04/12 17:00	QSIXYZ Q	1	4	1	4	Automatically exclude as no. of linked buses <2
2	01/04/12 18:00	QSIXYZ Q	1	4	0	4	Automatically exclude as no. of linked buses <2
2	01/04/12 19:00	QSIXYZ Q	1	2	2	0	
2	01/04/12 20:00	QSIXYZ Q	1	2	1	0	
2	01/04/12 21:00	QSIXYZ Q	1	2	0	4	Automatically exclude as no. of linked buses =0
2	01/04/12 22:00	QSIXYZ Q	1	1	1	0	
2	01/04/12 23:00	QSIXYZ Q	1	1	0	0	



## 4. Stage III: Manual Exclusion

---

### Overview

- 4.1. The automatic processes outlined in Sections 2 and 3 may, at the discretion of London Buses, be supplemented by additional manual exclusions where:
  - 4.1.1. A data error has not been removed satisfactorily by the imputation or automatic exclusion processes. For example, where removal of data recorded incorrectly at the next stop after buses have been diverted was incomplete.
  - 4.1.2. A result is, in the view of London Buses, unrepresentative of the service actually experienced by passengers.
  - 4.1.3. A wrong schedule (which may include link distance errors) has been loaded in iBus and, at the discretion of London Buses, the MTV mileage adjustment facility has been invoked. All QSI data for the duration of the incorrect schedule will be excluded. This applies to both **high and low frequency** routes
    - Where the correct schedule is loaded into iBus but an Operator operates to a different schedule in error, the MTV mileage adjustment facility will not be invoked and no exclusions will be undertaken.
  - 4.1.4. Where there is an identified system failure with the iBus data supply/collection system affective for 24 hours or more.
  - 4.1.5. There is an incorrect or corrupt Base Version release.
  - 4.1.6. Incorrect QSI locations being monitored.
- 4.2. Manual exclusions can be set up at any time (including in advance) but will usually be processed, as necessary, after the MTV Mileage Coding Finalisation, Data Imputation/Removal and Automatic Exclusion processes are complete.

### Unrepresentative Service

- 4.3. An 'unrepresentative' service or what is deemed an unrepresentative service will be determined by London Buses and would include:
  - 4.3.1. Major incidents that have an impact on the city-wide performance of the network.

- 4.3.2. Incidents, which resulted in a published instruction by London Buses to run a reduced service over a portion of the route due to a planned or unplanned event. A planned example would be a notice of event which explicitly states that the operator should run a reduced service over a particular portion of the route. An unplanned example of this would be where the operator has to divert part of the services away from QSI points, or there is a published explicit instruction on the part of London Buses to run a restricted service over an affected portion of route. Any data exclusions undertaken would be limited to only those locations within the affected section of route.
- 4.3.3. Where there is an identified technical issue with the respect to the data captured by the iBus system, which has been reported in the appropriate manner to London Buses and is verifiable. Examples would include diversion start/end location close to a QSI point causing partial data to be recorded, mistimings caused by corrupt data and trip false starts. Requests for data exclusions should be made via the 'Bus Operator Data Exclusion Request Process' (see Section 5 for details).
- 4.3.4. Unwarranted enhancements to services with the specific aim of improving QSI results (most likely at off peak times when spare drivers and vehicles may be available). This is not something that London Buses would accept given the potential negative impacts. These include exceeding stand capacity, unnecessary noise and emissions, and fuelling negative perceptions of "empty" buses given that the extra frequency is not required. These would outweigh the benefit to passengers who would gain from reduced headways. Where such undesired practices become apparent, London Buses reserves the right to exclude all affected data.
- 4.4. It should be emphasised London Buses expect Bus Operators to regard ad-hoc day-to-day incidents as 'normal' under the traffic and event conditions that occur when operating in the environment of a major city.
- 4.5. In all cases London Buses shall determine the reason and determine whether the data warrants exclusion based on the available information. Criteria used in this assessment will include:
  - 4.5.1. The route/direction(s)/location(s) and/or hourly time period(s) has been identified and agreed by London Buses as having an issue.
  - 4.5.2. The TSG reference number with respect to the problem. If London Buses consider a technical issue resolved, then exclusion will not be undertaken.
  - 4.5.3. A TfL published reference number and source of this information.
  - 4.5.4. Any other published TfL sourced data confirming the issue.
- 4.6. In all cases the Bus Operator shall restrict issues raised to published TfL/London Buses data. In all cases the operator will identify the source of the published reference numbers with respect to any issues raised.

- 4.7. Data will not be excluded where the resultant data errors have arisen due to failure on the part of the Operator to follow the correct processes. Examples of this include:
  - 4.7.1. Failure of drivers or staff to log into or use the apparatus correctly.
  - 4.7.2. Failure of staff to use the correct mileage cause codes which results in incorrect QSI data.
  - 4.7.3. Failure of the operator to follow the published instructions of London Buses. An example of this would be a failure to follow the instructed diversion, which results in data being captured at QSI survey locations.

## **System Failure**

- 4.8. Where there is an identified system failure with the iBus data collection system affecting collection for 24 hours or more, London Buses shall determine the impact and severity of the failure using published TfL sourced information. Examples of this would be a failure of computer hardware.
- 4.9. Data may be excluded where the following criteria are met:
  - 4.9.1. Outages of a concurrent period of 24 hours or more.
  - 4.9.2. Failures of the iBus telecommunications or data network infrastructure provided by London Buses or its contractors, for a concurrent period of 24 hours or more.
  - 4.9.3. iBus system failures which result in a 100% failure rate of the Vicos-Lio controlling screen workstations for a garage or identified controlling centre, which occur as a result of failure of the provided iBus System.
- 4.10. In the event of any exclusion submissions, the operator will provide printed or electronic supporting information as to the steps taken to reinstate service control outside of the iBus system for periods of 24 hours or more.
  - 4.10.1. The operator will clearly demonstrate in the supporting evidence that additional service control was implemented. This information will be supplied by the operator on a daily and route-by-route basis.
- 4.11. London Buses will decide as a result of the supporting evidence provided as to whether exclusion is warranted.
- 4.12. Data will not be excluded under the following criteria:
  - 4.12.1. Vicos-Lio outages of 24 hours or less for any reason.
  - 4.12.2. Failures of individual Vicos-Lio controlling workstations at control centres.
  - 4.12.3. Failures as a result of events within the 'control' of the operators, such as power failures to control rooms, failures of the operator's communication systems or other events such as damage to buildings which prevent operation of the Vicos-Lio controlling screens.

- 4.12.4. Failures of individual hardware equipment on individual buses.
- 4.12.5. Failure to adhere to the correct sign in/sign out process on individual buses, which results in the trip failing to be registered.
- 4.12.6. Failures of individual buses to record at individual QSI points.

## 5. Bus Operator Data Exclusion Request Process

---

- 5.1. Operator requests for Data Exclusions should not be made until London Buses has confirmed that provisional period QSI results have been completed. This is important to prevent unnecessary exclusion requests as prior to such notification being issued the data is still under review and consequently QSI results could change. After this, an Operator shall be permitted to submit requests, according to contractual deadlines, for the review of specific data for which they believe continues to contain errors following completion of the data cleansing processes.
- 5.2. Following investigation London Buses may decide, at its absolute discretion, to exclude some or all of the data specified in the request. Data exclusions will not be given for events deemed to be accounted for by the Automatic Exclusion process. See Section 3 (item 3.2) for details.
- 5.3. London Buses closely monitors the level of requests being submitted and Operators should take care to submit only requests that may qualify for exclusion under the specified guidelines (see Section 4 for details). Should an Operator continue to submit excessive levels of non-qualifying requests, London Buses reserve the right to suspend all requests from that Operator for a minimum of 1 period.
- 5.4. The operator should provide as much information as possible on any issue being submitted for review. Submissions and any supporting evidence should be provided electronically, via the Data Query Submission Form with scanned electronic versions of any supporting evidence attached. Operators have been provided with the Data Query Submission Form spreadsheet and the following table provides a guide to the information required:

Category	Comments
Operator	e.g. 'XX'
Garage	e.g. 'YZ'
Route	e.g. '1234'
The start date of the incident	e.g. 23/12/2012
The start time of the incident	e.g. 23:00
The end date of the incident	e.g. 24/12/2012
The end time of the incident	e.g. 01:00
QSI points	The QSI Point(s) being requested (e.g. ABCDE Q)
Direction	The direction(s) : 1, 2 or both
Incident reference number(s)	TSG TechServices Helpdesk reference number Or TfL sourced published reference number The source of the reference number must be stated.
Explanation of issue	The reason why the data is being proposed for review.

## 6. Data and Process Integrity

---

- 6.1. London Buses will undertake whatever actions it decides are required to maintain the integrity of the iBus data and QSI monitoring system. This may include but not be limited to:
  - 6.1.1. Regular data sampling and checks for variations in the volume, type, time and location of data exclusions.
  - 6.1.2. Investigations and process checks.
  - 6.1.3. Audits of QSI results, mileage coding and exclusions at operator, garage and route level.
- 6.2. London Buses may decline to exclude data, where there is evidence of miscoding of mileage data. Examples of this would be incorrect mileage cause codes being used, or discrepancies between evidence provided by the operator.
- 6.3. London Buses may request information from an operator, such as log cards or ticket module data, to facilitate the above scrutiny. Operators are required to provide this data, or to make facilities available for London Buses staff to review this data as and when required. Any request for data should be supplied no later than 10 working days after the request is made.

## 7. Data Exclusion Codes

---

- 7.1. Each route/direction/QSI point/hour can be assigned one or more exclusion types within iBus. For example, an hour marked as excluded through imputation may also qualify for exclusion via the automatic process. In order to differentiate between imputation, automatic and manual exclusions within the database, each potential combination has been assigned a unique code.
- 7.2. Where an hour is excluded due to a combination of factors, there remains a 'principle exclusion' reason. For example, an hour marked as excluded through imputation may also be subject to a manual exclusion but the principle exclusion will be 'imputation' since the imputation outcome would happen regardless of any manual exclusion being applied. Similarly, an hour marked as excluded through imputation which also fulfils automatic exclusion criteria would have a principle exclusion of 'imputation' since the imputation process takes place first.
- 7.3. The 'exclusion codes' and 'principle exclusion' reasons are described in the table below. As imputation exclusions are only applicable for High Frequency routes, codes 2, 3, 6 and 7 are not used for Low Frequency routes. The shared combinations have consistent codes applied for both route frequency types.

Exclusion Code	Imputation	Automatic	Manual	Principle Exclusion	Route Frequency
0	No	No	No	N/A	HF & LF
1	No	No	Yes	Manual	HF & LF
2	Yes	No	No	Imputation	HF only
3	Yes	No	Yes	Imputation	HF only
4	No	Yes	No	Automatic	HF & LF
5	No	Yes	Yes	Automatic	HF & LF
6	Yes	Yes	No	Imputation	HF only
7	Yes	Yes	Yes	Imputation	HF only

- 7.4. The codes are included for information only and whichever code (1 to 7) is assigned to an hourly total, the impact on QSI calculations is the same. The individual codes are not currently distinguished in any of the Hyperion QSI Performance Reports available to Operators although they may appear elsewhere and might be made available in future reporting enhancements.

## 8. Hyperion Performance Reports

---

### Overview

- 8.1. Information on data excluded as a result of the imputation and data cleansing processes is presented in the Hyperion QSI Performance Reports in various ways.
- 8.2. The timing of when reports are run is an important factor as the QSI results may change following completion of each process.

### Unfinalised Data

- 8.3. With the exception of R210 'Trip Observation Detail', which runs only against finalised data, the QSI Performance Reports can be run for dates as recent as current date minus 2.
- 8.4. This means that reports can be run before the weekly automatic data finalisation and data cleansing processes have been completed. Users should be aware that the QSI results could change after each automatic process has run.
  - 8.4.1. Where finalisation has not taken place for any of the data displayed in a report output the report header will state '**Results contain unfinalised data**' in the top right corner of each page.
- 8.5. In certain circumstances manual exclusions may be undertaken AFTER finalisation. When this occurs, London Buses will notify Operators by email of any changes to their QSI results.

### 100% Data Excluded

- 8.6. Where a report data table contains a row where all data has been excluded, data will be shown as follows:
  - Expected buses will be shown as '0'.
  - Observed Buses / % On AVL will be shown as '0'.
  - Performance measures (for example EWT or % On-time) will be shown as 'NA'.



## **R210 'Trip Observation Detail'**

- 8.7. This report shows individual observation times, which will be marked as 'Exc' or 'Dis' where they have been excluded (for any reason) or removed (LF only) from the QSI calculations.
- 8.8. Imputed 'observation' times for High Frequency routes are also show on this report and are distinguished by *italic* text.